Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2021) 29: 2215 – 2231

doi:10.3906/elk-2010-91

Clustering ensemble selection based on the extended Jaccard measure

Hajar KHALILI[1], **Mohsen RABBANI**[2,*], **Ebrahim AKBARI**[1]
[1]Department of Computer Engineering, Sari Branch, Islamic Azad University, Sari, Iran
[2]Department of Applied Mathematics, Sari Branch, Islamic Azad University, Sari, Iran

Received: 20.10.2020 • **Accepted/Published Online:** 16.03.2021 • **Final Version:** 26.07.2021

Abstract: Clustering ensemble selection has shown high efficiency in the improvement of the quality of clustering solutions. This technique comprises two important metrics: diversity and quality. It has been empirically proved that ensembles of higher effectiveness can be achieved through taking into consideration the diversity and quality simultaneously. However, the relationships between these two metrics in base clusterings have remained uncertain. This paper suggests a new hierarchical selection algorithm using a diversity/quality measure based on the Jaccard similarity measure. In the proposed algorithm, the selection of the subsets of the clustering partitions is done based on their diversity measures. The proposed diversity measure (in two types of pair-wise diversity and hybrid diversity) is applied to the proposed algorithm. Hypergraph-partitioning algorithm (HGPA), cluster-based similarity partition algorithm (CSPA), and meta-clustering algorithm (MCLA) were used to obtain the consensus solution and cluster ensemble selection results with a hierarchical method. The experimental results on 14 datasets showed that selecting a subset of base clusterings using the proposed algorithm led to more accurate results compared to those of the full ensemble. The effectiveness and robustness of the proposed algorithm were demonstrated in comparison with the full ensemble. The comparative results showed that the proposed method by new diversity measure outperformed the full ensemble.

Key words: Cluster ensemble selection, diversity, quality, extended Jaccard measure

1. Introduction

Data clustering or unsupervised learning is a fundamental conceptual principle in data mining. Data clustering is also recognized in the literature as cluster analysis, whose main objective is to partition a set of unlabeled objects into a number of homogeneous groups or clusters [1, 2]. With the rapid advancement of clustering technology, cluster analysis plays an important role in a variety of areas, including pattern recognition, social network analysis, document clustering, bioinformatics, and image segmentation, etc [3, 4]. It is difficult to find a clustering algorithm that can be applied to all data sets and to mark out an appropriate algorithm for a certain dataset. Therefore, different clustering algorithms have been proposed and improved in the literature [5]. To solve this problem, authors have suggested the concept of clustering ensemble [6, 7].

The clustering ensemble (CE) aggregates several clustering results for the aim of obtaining final clusters and attempts to enhance the accuracy and efficiency of clustering results. Generally, the clustering ensemble involves two steps. The first step is about creating a diverse set of base clusterings, which should be different from each other because a high level of diversity between the base clusterings can potentially help to improve the performance of the ensemble solution [8, 9]. The second step is the use of a consensus function that is

*Correspondence: mrabbani@iausari.ac.ir

an algorithm that integrates all the clusterings obtained at the first step to achieve final clusters [10, 11]. Traditionally, all of the base clusterings are combined together by a consensus function. Although objects in the clustering problem are unlabeled, some results may be unreliable within an extensive clusterings library. Therefore, it cannot be expected that all achieved clustering results be beneficial to the final solution of consensus clustering [12, 13].

Recently, it has been proved that, with the use of a subset of clustering members, better clustering results can be achieved [14]. This approach is named clustering ensemble selection (CES). The selection strategy aims to select better clusterings from among base clusterings. The main idea of cluster ensemble selection is selecting a diverse subset of base clusterings for the formation of a smaller-sized cluster ensemble performing better than the set of all available ensemble members [12]. Figure 1 demonstrates the clustering ensemble selection evolution.

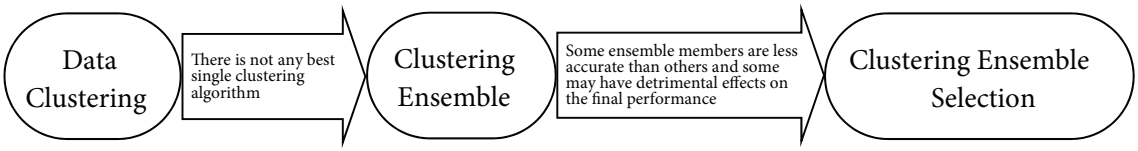

Figure 1. Clustering ensemble selection evolution.

Diversity and quality have been identified as two factors of high importance, which need to be considered to select the basic clustering and affect ensemble performance. Diversity is critical to the success of ensemble clustering, and base clusterings of high quality affect positively the final solution performance. [15] showed that the use of both diversity and quality in CES results in the enhancement of the final results in comparison with full ensembles. The relationship between diversity and quality is uncertain. To enhance the quality of the ensemble, diversity is increased by removing the redundant base partitions [16].

The present study attempts to propose a novel diversity criterion applicable to CES. Initially, in this method, a pair-wise (or hybrid) diversity matrix comprising all accessible ensemble members is created. Then, a hierarchical clustering algorithm based on the single-link method is applied using a diversity matrix. A subset of clustering members is selected based on the quality measure. Finally, a consensus function is used to obtain the consensus solution. In the following, the most important contributions of the present research are presented:

(i) Proposing a diversity/quality measure.
(ii) Applying hierarchical cluster ensemble selection on the basis of the diversity criterion.

The remaining parts of this paper are presented as follows. An overview of related work is discussed in Section 2. Section 3 presents different diversity and quality criteria. Section 4 introduces the method proposed in this paper. Section 5 explains the experiments and the obtained results. Finally, the paper is concluded in section 6.

2. Related Work

In general, the objective of clustering ensemble approaches is the improvement of the robustness and quality of clustering results. Many approaches exist in the literature aiming to offer effective solutions to clustering ensemble problems [17, 18]. Ensemble clustering comprises two stages: diversity (the creation of multiple clusterings) and consensus function (the aggregation of multiple clusterings). Most of the previously proposed methods have used all base clusterings for the final clustering; however, a number of scholars in recent years, have

developed the diversity selection to enhance the ensemble performance. Cluster ensemble selection approaches have been presented to enhance the way the ensemble performs [19, 20]. The main problem in the clustering ensemble selection is how to evaluate each cluster. The literature consists of different quality and diversity measures implemented to ensemble members [21, 22]. The basis of these measures is the match index between the two partitions. Normalized mutual information (NMI) [6] and adjusted Rand index (ARI) [23] are two diversity measures that have been employed by many researchers. These measures are also used to measure the quality between two partitions. For example, Zhong and Ghosh [24] applied the NMI to evaluating clusters, while Kandylas et al. [25] applied it to community knowledge analysis. In another study, Hadjitodorov et al. [14] used the ARI diversity measure to select ensemble members. Lu et al. [21] suggested a diversity criterion based on covariance. Alizadeh et al. [26] proposed a CES method in which clusters were selected based on diversity and quality measures.

Hadjitodorov et al. [14] formed a large number of base clusterings. Then, the ensemble members with the median diversity were selected for the purpose of generating the final solution. Moreover,Fern and Lin [15] examined a variety of heuristics for choosing subsets of ensemble members. They designed a number of methods for ensemble selection on the basis of quality and diversity and used the NMI measure to select clusterings. Their method, first, clusters all base clusterings using a spectral clustering algorithm; afterward, it selects a single solution from each of the available clusters in order to build the final ensemble. Azimi and Fern [12] proposed an adaptive clustering ensemble method that selected a subset of clusterings adaptively. They used the NMI criterion to evaluate all base clusterings and divided them into *stable* and *non-stable* members. Jia et al. [27] developed a clustering ensemble method, namely selective spectral clustering ensemble (SELSCE), based on the bagging technique. They generalized the algorithm proposed in [12] and used multiple approaches to produce the components of the ensemble system. Hong et al. [13] suggested a CES method based on the resampling method. Parvin et al. [19] proposed a method for clustering the available data using weighted features. In this method, the data variance through every feature is calculated. Then, the feature whose variance is higher participates in combination. Naldi et al. [22] presented various relative cluster validity indexes on the basis of quality and diversity for the purpose of clustering selection. In addition, they studied how diversity affects the partitions employed in the ensemble.

Alizadeh et al. [2, 26] suggested the selection of a subset of clusters to form final clustering instead of selecting a subset of base clusterings. Akbari et al. [28] introduced the hierarchical cluster ensemble selection (HCES) method. The HCES method finds the subset of cluster members on the basis of diversity and quality. In HCES, all available clusterings are clustered in a hierarchical way. Yousefnezhad et al. [29] suggested a new strategy for applying graph-based modeling to evaluate the independence of the basic clustering algorithm when selecting the cluster ensemble. Independence can be assessed well with the use of a novel modeling language, called clustering algorithms independence language (CAIL), whereas the assessment of diversity can be done using the APMM criterion that is presented in [2]. Yang et al. [30] developed a unified framework for the aim of solving the constraint-based clustering ensemble selection problem. They simultaneously considered the quality and diversity under the prior background information. The experimental results demonstrated that their method (AQD2) performed best because of its promising quality and consistency with a satisfactory diversity. Shi et al. [31] proposed the transfer cluster ensemble selection (TCES) algorithm for an adaptive selection of clustering members on the basis of the relationships between quality and diversity. Additionally, they introduced a transfer CE framework (TCE-TCES) on the basis of TCES for the aim of achieving more acceptable clustering

results. The results of their experiments confirmed TCE-TCES as a successful model in finding a better trade-off between quality and diversity, and also it was found capable of obtaining more desirable clustering results.

Wang and Liu [32] proposed a selective clustering ensemble approach. In this algorithm, partitions are selected by hybrid multi-modal metrics. The results of the experiments conducted in their study indicated that their algorithm is capable of selecting partitions concerning both diversity and quality and also getting robust clustering results. Li et al. [33] proposed a selective clustering ensemble framework that takes into account the difference between the demand in the ensemble selection stage and the demand in the ensemble integration stage. Moreover, they introduced an innovative measure, called SME, and confirmed that due to some certain properties, SME has the required capacity for measuring the quality of each cluster in the ensemble. Abbasi et al. [34] developed a method applicable to the selection of a subset of clusters with higher effectiveness. They suggested a criterion based on NMI, called edited NMI (ENMI), for cluster evaluation, and experimentally showed that, in general, the ENMI criterion was slightly more successful than MAX and NMI criteria. Additionally, they proposed a method capable of building the co-association matrix, namely, extended evidence accumulation clustering (EEAC). Their experimental results demonstrated that the most effective option for consensus function is EEAC with the average-linkage algorithm. Liang et al. [35] proposed two decision-making methods based on information entropy to select clusterings. They showed that both methods can improve ensemble clustering methods.

However, there is still a challenge to find the relationship between diversity and quality and their influence on the ensemble performance. Some consensus functions, e.g., cluster-based similarity partition algorithm (CSPA) and meta-clustering algorithm (MCLA), have more accurate ensemble solutions, while base clusterings have moderate diversity. Though, some functions such as hypergraph-partitioning algorithm (HGPA) require more diverse base clusterings [28]. Accordingly, the present paper suggested a new hierarchical selection algorithm using a diversity/quality measure.

3. Diversity and quality measures

The literature consists of different diversity measures; the majority of them are based on matching the labels obtained from two partitions. Two partitions are assumed diverse if their labels are not matched completely. NMI and ARI are two diversity criteria commonly used in the literature; among these two, NMI is used in our experiments.

Suppose that $X = \{x_1, x_2, \ldots, x_n\}$ is a set of n data points. In clustering ensemble problem, several clusterings are created. Assume that $H = \{h^1, h^2, \ldots, h^L\}$ is a set of L clusterings. A clustering h^i of X is k_i groups of data points as $h^i = \{c_1^i, c_2^i, \ldots, c_{k_i}^i\}$, where c_p^i is a cluster of clustering h^i, $i = 1, 2, \ldots, L$; $p = 1, 2, \ldots, k_i$ and k_i is the number of clusters in clustering h^i. Diversity measure can be defined based on quality measure as follows:

$$d(h^i, h^j) = 1 - \delta(h^i, h^j) \tag{1}$$

where $d(h^i, h^j)$ stands for diversity between h^i and h^j, and δ is a quality criteria such as ARI or NMI in which the value of δ is between 0 and 1. Suppose that h^* is the result of a consensus function.

$$h^* = \Phi(H, k) \tag{2}$$

where Φ is a consensus function on the basis of a quality measure, and k stands for the number of clusters. h^* is

a clustering that has maximal average mutual information with all individual labelings h^i [6] and is formalized as:

$$h^* = arg\max_{\hat{h}} \sum_{i=1}^{r} NMI(\hat{h}, h^i) \tag{3}$$

where $\hat{h}$ goes through all possible clusterings. Based on h^*, diversity between h^* and h^i is defined as:

$$d(h^*, h^i) = 1 - \delta(h^*, h^i) \tag{4}$$

We can divide diversity measures into pair-wise, non-pair-wise, and hybrid. The first item is calculated using base clusterings. Non-pair-wise diversity is calculated based on h^*, which is obtained by a consensus function, e.g., CSPA, HGPA, or MCLA (Eq. 2). Hybrid diversity is a combination of pair-wise diversity and non-pair-wise diversity. On the other hand, hybrid diversity is a combination of h^* and base clusterings.

3.1. Pair-wise diversity

Pair-wise diversity is a type of internal diversity. Each clustering in pair-wise diversity is selected as a reference class label, and other clusterings are compared with the reference label. Pair-wise diversity for each h^i is defined by Eq. (5),

$$dp(h^i) = \frac{1}{L-1} \sum_{j=1, j\neq i}^{L} d(h^i, h^j) \tag{5}$$

and pairwise diversity of H is defined by Eq. (6),

$$Dp(H) = \frac{1}{L} \sum_{i=1}^{L} dp(h^i) \tag{6}$$

where $Dp(H)$ is the average of pair-wise diversity, and H is a set of L clusterings.

3.2. Non-pair-wise diversity

Non-pair-wise diversity is another type of internal diversity in which, for each clustering, h^i is calculated by Eq. (7). Non-pair-wise diversity is diversity between h^i and h^*,

$$dnp(h^i) = d(h^*, h^i) \tag{7}$$

where h^* is obtained by a consensus function (Eq. 2). Non-pair-wise diversity of H is obtained using Eq. (8).

$$Dnp(H) = \frac{1}{L} \sum_{i=1}^{L} dnp(h^i) \tag{8}$$

where H is a set of L clusterings, and $Dnp(H)$ is the average of non-pair-wise diversity.

3.3. Hybrid diversity

Hybrid diversity measure is applied to the computation of the distance between diversity of h^i and h^j. Hybrid diversity is a combination of pair-wise diversity and non-pair-wise diversity. Hybrid diversity measure between h^i and h^j is calculated using Eq. (9).

$$Hd(h^i, h^j) = |dnp(h^i) - dnp(h^j)| \tag{9}$$

Hybrid diversity for each h^i is defined by Eq. (10),

$$Hd(h^i) = \frac{1}{L-1} \sum_{j=1, j\neq i}^{L-1} Hd(h^i, h^j) \tag{10}$$

where $Hd(h^i, h^j)$ is obtained by Eq. (9). The average of hybrid diversity is calculated with the use of Eq. (11).

$$Hd(H) = \frac{1}{L} \sum_{i=1}^{L} Hd(h^i) \tag{11}$$

4. Proposed method

In the present article, a novel criterion is suggested on the basis of the Jaccard measure, which is called extended Jaccard (EJ). This measure is calculated by Eq. (12).

$$EJ(h^a, h^b) = \frac{1}{2} \left[\frac{1}{k_a} \sum_{i=1}^{k_a} max \left\{ Jaccard(c_i^a, c_j^b) \right\}_{j=1}^{k_b} + \frac{1}{k_b} \sum_{j=1}^{k_b} max \left\{ Jaccard(c_i^a, c_j^b) \right\}_{i=1}^{k_a} \right] \tag{12}$$

$$Jaccard(a, b) = \frac{a.b}{||a||^2 + ||b||^2 - a.b} \tag{13}$$

In Eq. (12), h^a and h^b are two clusterings of H and $h^a = \{c_1^a, c_2^a, \ldots, c_{k_a}^a\}$, $h^b = \{c_1^b, c_2^b, \ldots, c_{k_b}^b\}$, k_a and k_b are the numbers of clusters in clusterings h^a and h^b, respectively. The Jaccard measure is generally known as a similarity measure that is defined between two clusters and is calculated by Eq. (13). The value of the Jaccard similarity measure is between 0 and 1. When two clusters are completely mismatched, the value of Jaccard is 0, while the value of 1 shows that two clusters are matched completely.

As an example, let $X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$. Two clusterings are presented on dataset X by set representation, which are $h^1 = \{\{x_1, x_2\}, \{x_3, x_4\}, \{x_5, x_6\}\}$, $h^2 = \{\{x_1, x_2, x_3\}, \{x_4, x_5, x_6\}\}$. Clustering h^1 contains three clusters: $C_1^1 = \{x_1, x_2\}$, $C_2^1 = \{x_3, x_4\}$ and $C_3^1 = \{x_5, x_6\}$; whereas clustering h^2 contains two clusters: $C_1^2 = \{x_1, x_2, x_3\}$ and $C_2^2 = \{x_4, x_5, x_6\}$. The values of the Jaccard similarity measure between clusters of h^1 and h^2 are shown in Table 1. The quality between clusterings h^1 and h^2 are: $NMI(h^1, h^2) = 0.5158$, $ARI(h^1, h^2) = 0.2424$, $EJ(h^1, h^2) = 0.5972$. The value of EJ measure is between 0 and 1 similar to NMI. If two clusterings are the same, e.g., $h^1 = \{\{x_1, x_2\}, \{x_3, x_4\}, \{x_5, x_6\}\}$ and $h^2 = \{\{x_3, x_4\}, \{x_5, x_6\}, \{x_1, x_2\}\}$; then the quality between these clusterings are: $NMI(h^1, h^2) = 1.0$, $ARI(h^1, h^2) = 1.0$, $EJ(h^1, h^2) = 1.0$.

Table 1. The values of Jaccard measure between clusters of h^1 and h^2.

	C_1^2	C_2^2
C_1^1	2/3	0
C_2^1	1/4	1/4
C_3^1	0	2/3

Algorithm 1 : Hierarchical Algorithm with Single Link

1. Calculate the distance from each cluster to all other clusters. Note that at first, each cluster includes the primary points.
2. Identify a pair of clusters, (r, s), with the shortest distance.
3. Merge clusters r and s and recompute the distances between clusters.
4. Repeat Step 2 and 3 until all clusters are merged in one cluster.

The proposed algorithm selects a subset of base clusterings with more diversity considering pair-wise diversity and hybrid diversity, in which δ is EJ. These two diversity measures were used in our experiments, and their effects on the quality of final clustering were compared with those of the full ensemble. Initially, all clusterings $H = \{h^1, h^2, \ldots, h^L\}$ are clustered into k clusters as $\Pi = \{C_1, C_2, \ldots, C_k\}$, where $C_i = \{h_1^{c_i}, h_2^{c_i}, \ldots, h_{r_i}^{c_i}\}$, $i = 1, 2, \ldots, k$, $\sum_{i=1}^{k} r_i = L$, r_i stands for the number of clusterings of cluster C_i, and L is the number of base clusterings. Akbari et al. [28] employed the hierarchical clustering algorithm using single-link, complete link, and average-link methods. They showed that the single-link outperformed the others. Therefore, we use the single-link method for the purpose of this study. Algorithm 1 shows the single-link procedure [36]. At each step, a diversity matrix based on M^0 is calculated and used by the single link method (Eq. 14).

$$M^0 = \begin{bmatrix} m_{11} & m_{12} & \ldots & m_{1s} \\ m_{21} & m_{22} & \ldots & m_{2s} \\ \vdots & \vdots & & \vdots \\ m_{s1} & m_{s2} & \ldots & m_{ss} \end{bmatrix} \tag{14}$$

where m_{ij} is the value of pair-wise diversity and hybrid diversity, which is calculated using Eq. (15) and Eq. (16), respectively.

$$m_{ij} = \begin{cases} 0 & \text{if } i = j, \\ 1 - EJ(h^i, h^j) & \text{if } i \neq j. \end{cases} \tag{15}$$

$$m_{ij} = \begin{cases} 0 & \text{if } i = j, \\ EJ(h^*, h^i) - EJ(h^*, h^j) & \text{if } i \neq j. \end{cases} \tag{16}$$

where h^* is obtained by a consensus function (Eq. 2). At the first level, all clusters (L clusters) are combined by a consensus function, e.g., CSPA, HGPA, or MCLA, and h^* is obtained. At the second level, the matrix M^1 is constructed and the hierarchical algorithm with the single-link method is used to build new clusters. At

this level, the number of clusters is $L-1$. This procedure continues the $L-1$ step until two clusters remain. At each step of the single-link algorithm, the number of clusters is shown in Fig. 2.

Algorithm 2 : Proposed Algorithm

1. Generating ensemble members.
2. Applying a consensus function to obtain consensus solution h^*.
3. Constructing pair-wise (or hybrid) diversity matrix based on Eq. (15) or Eq. (16).
4. Applying the single-link algorithm using diversity matrix.
5. Selecting one clustering, h_i^* , with the highest NMI value from each group.
6. Applying a consensus function on the selected subset $\{h_1^*, h_2^*, \ldots, h_k^*\}$ and produce h_c^*.
7. Selecting the best solution based on ensemble qualities.

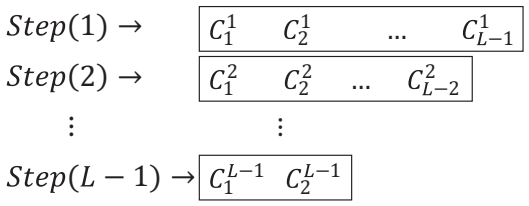

Figure 2. Number of clusters at each step of the proposed algorithm.

At each step, a diversity matrix, M^s, where $s = 1, 2, \ldots, L-1$, which is based on M^0, is used. The diversity matrix M^1 is produced based on $C_1^1, C_2^1, \ldots, C_{L-1}^1$, while the diversity matrix M^2 is produced based on $C_1^2, C_2^2, \ldots, C_{L-2}^2$, and finally, the diversity matrix M^{L-1} is produced based on C_1^{L-1}, C_2^{L-1}. Each C_j^i is a cluster of clusterings and contains some clusterings like $h_1^i, h_2^i, \ldots, h_{r_i}^i$. In each cluster, one sample (clustering) is selected using Eq. (17), which is of high quality. As a result, at each level, we have some selected members like $h_1^*, h_2^*, \ldots, h_k^*$; this is the cluster ensemble selection phase of the proposed method.

$$h_i^* = arg\max \left\{ \frac{1}{r_i - 1} \sum_{j=1, j\neq l}^{r_i} NMI(h_l^{c_i}, h_j^{c_i}) \right\}_{l=1}^{r_i}, i = 1, \ldots, k \tag{17}$$

At step k, where there are k clusters of clusterings in which $\Pi = \{C_1, C_2, \ldots, C_k\}$, where, $C_1 = \{h_1^{c_1}, h_2^{c_1}, \ldots, h_{r_1}^{c_1}\}$, $C_2 = \{h_1^{c_2}, h_2^{c_2}, \ldots, h_{r_2}^{c_2}\}$, and, $C_k = \{h_1^{c_k}, h_2^{c_k}, \ldots, h_{r_k}^{c_k}\}$, k members, $h_1^*, h_2^*, \ldots, h_k^*$ are selected using Eq. (17); each member is chosen from a different cluster. Then, a consensus function is used to combine these selected members and create h_c^*. Afterward, h_c^* and h^* are compared to each other. Figure 3 shows the proposed method at step k, and Algorithm 2 is a pseudo code of the proposed algorithm.

Algorithm 2 begins by generating initial clusterings (line 1) and then applying a consensus function, e.g., CSPA, HGPA, or MCLA, and obtains a consensus solution (line 2). Then, a diversity matrix is calculated using Eq. (15) or Eq. (16) (line 3). The algorithm continues by applying the single-link algorithm (line 4). Next, in each cluster, one sample is selected using Eq. (17) (line 5). Finally, the selected samples are combined with a consensus function (line 6).

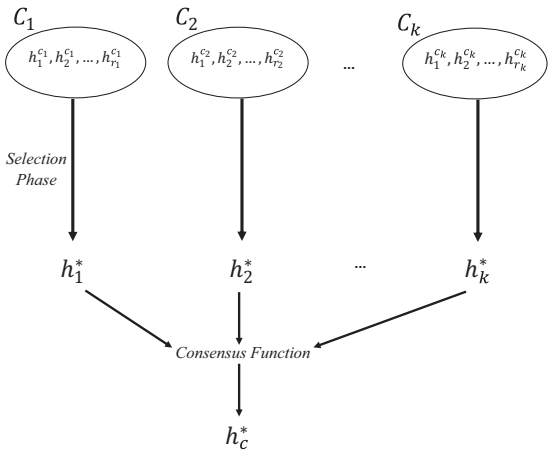

Figure 3. Main stages of the proposed algorithm at step k.

5. Experimental results

This section evaluates the way the proposed algorithm performs its tasks defined and the proposed criterion through applying them to real and artificial datasets. The quality of the clustering results was assessed using the NMI and EJ criteria. Note that if the values of NMI and EJ criteria are 0, two clusterings are completely different, while they are similar if the values are 1. Thus, the performance of the proposed method using the NMI and EJ criteria was compared with that of the full ensemble. Three graph-based consensus functions, i.e., CSPA, HGPA, and MCLA, were employed in the experiments to find the final solution. This study makes use of the NMI criterion for the aim of evaluating the final solution.

5.1. Datasets

The performance of the method proposed in this paper was evaluated over eleven standard datasets and three artificial datasets. The datasets considered in this study are presented in Table 2 where n is the number of instances, d denotes the number of attributes of a dataset, and k stands for the number of classes in a dataset. The samples of these datasets range between 47 and 6435. The attribute number of these data ranges from 2 to 36. The cluster numbers of these data are between 2 and 10.

5.2. Generating ensemble members

Since in the literature, k-means has been reported as one of the best options to generate base clusterings, in the present paper, the k-means algorithm was used with different k values and initializations for the purpose of producing ensemble members (base clusterings). In our experiments, different k, and for each k, different runs of the k-means algorithm were chosen to produce diversity. The k-means algorithm generates M ensemble members, in which $M = r * s$ where k-means is run r times for various k, and for each k, it is run s times.

$$k = \begin{cases} [\sqrt{n} - r : \sqrt{n}] & \text{if } \sqrt{n} > r, \\ [\frac{\sqrt{n}}{2} : \frac{\sqrt{n}}{2} + r] & \text{if} \sqrt{n} \leq r. \end{cases} \tag{18}$$

Finally, a set of M base clusterings is generated. In the literature, the value of k is considered in the range of $[2, \sqrt{n}]$ for generating base clusterings [18, 22]. This paper suggested that the value of k can be determined by Eq. (18).

Table 2. Datasets used in experiments.

No.	Dataset	n	d	k
1	Jain	373	2	2
2	Path bass	300	2	2
3	Aggregation	788	2	7
4	Soybean(small)	47	16	4
5	Breast-tissue	106	9	6
6	Iris	150	4	3
7	Wine	178	13	3
8	Seeds	210	7	3
9	Glass	214	10	7
10	Ecoli	336	7	8
11	Breast-cancer	699	9	2
12	Yeast	1484	8	10
13	Segmentation	2310	19	7
14	Satimage	6435	36	7

The suggested method generated more diverse partitions. The k-means generated 100 primary partitions. In the first method, the value of k was in the range of $[2, \sqrt{n}]$, and, in the second method, the value of k was chosen based on Eq. (18). For example, to generate $M = 100$ base clusterings for the Iris dataset, while the size of this dataset was $n = 150$ and $\sqrt{150} \approx 12$, in the first method, the value of k was chosen from 2 to 12, and, in the second method, the values of r and s were set to 20 and 5. Since $\sqrt{n} \leq r$, the value of k was selected from the values ranging between 6 and 26.

Three consensus functions (i.e., CSPA, MCLA, and HGPA) were used to obtain a consensus solution. Since the labeling for datasets was available, the NMI criterion was calculated for the evaluation of the final solution. Tables 3 and 4 show the quality and diversity of the results obtained by consensus functions. In Table 4, the qualities and diversities of base clusterings are better than the qualities and diversities shown in table 3. For the soybean (small), breast-tissue, iris, seeds, breast-cancer, yeast, segmentation, and Satimage datasets, the proposed method achieved higher quality in all of the three consensus functions. For the Iris, glass, and Ecoli datasets, the proposed method achieved higher quality in the CSPA and HGPA algorithms, whereas it achieved lower quality in the MCLA algorithm. In all datasets, the proposed method achieved higher diversity in both pair-wise and hybrid diversities. As a result, when we changed the value of k using Eq. (18), base clusterings with better diversity were generated. Furthermore, the second method generated better diversity than the first method. According to Tables 3 and 4, the impact of quality and diversity on consensus functions can be obtained. A higher diversity of base clusterings leads to a higher quality of the consensus solution, if there is the quality between base clusterings.

5.3. Test results based on diversity

In this section, different diversity measures were investigated through varying the ensemble size from 10 to 100. For each ensemble size, the proposed method was run ten times. Then, the average value was calculated as the final result. Figures 4, 5, and 6 compare the final solution diversity based on different diversity measures. The diversity measure of $[1 - NMI(h^i, h^j)]$ was shown with $d1 - diversity$, $[1 - EJ(h^i, h^j)]$ was shown with $d2 - diversity$, $[NMI(h^*, h^i) - NMI(h^*, h^j)]$ was shown with $d3 - diversity$, and $[EJ(h^*, h^i) - EJ(h^*, h^j)]$

Table 3. Clustering quality and diversity for different $k \in [2, \sqrt{n}]$.

Dataset	CSPA	MCLA	HGPA	Pair-wise diversity	Hybrid diversity
Soybean(small)	0.737	0.846	0.646	0.204	0.094
Breast-tissue	0.302	0.354	0.306	0.216	0.142
Iris	0.845	0.865	0.857	0.299	0.080
Wine	0.389	0.394	0.397	0.250	0.068
Seeds	0.676	0.688	0.668	0.288	0.066
Glass	0.354	0.381	0.309	0.287	0.056
Ecoli	0.507	0.591	0.488	0.280	0.040
Breast-cancer	0.466	0.753	0.471	0.282	0.034
Yeast	0.213	0.239	0.184	0.309	0.024
Segmentation	0.476	0.540	0.325	0.164	0.011
Satimage	0.462	0.472	0.327	0.185	0.047

Table 4. Clustering quality and diversity for different k using proposed method.

Dataset	CSPA	MCLA	HGPA	Pair-wise diversity	Hybrid diversity
Soybean(small)	0.754	0.862	0.679	0.288	0.185
Breast-tissue	0.315	0.368	0.338	0.431	0.285
Iris	0.919	0.919	0.901	0.305	0.103
Wine	0.346	0.395	0.425	0.302	0.079
Seeds	0.684	0.695	0.702	0.343	0.083
Glass	0.367	0.345	0.321	0.366	0.118
Ecoli	0.513	0.587	0.500	0.346	0.083
Breast-cancer	0.485	0.812	0.485	0.360	0.062
Yeast	0.227	0.243	0.199	0.432	0.063
Segmentation	0.488	0.549	0.334	0.335	0.068
Satimage	0.551	0.524	0.473	0.294	0.063

was shown with $d4 - diversity$. $d1 - diversity$ and $d2 - diversity$ are pair-wise diversities and need to be compared with each other. $d3 - diversity$ and $d4 - diversity$ are hybrid diversities and need to be compared with each other, too. For all datasets, the proposed method, which was based on $d2$ diversity measure, obtained clusters with more diversity compared to the clusters obtained based on $d1$ diversity measure (Figs.(4), (5) and (6)). Additionally, the proposed method based on the $d4$ diversity measure obtained more diverse clusters than the obtained clusters using the $d3$ diversity measure.

Typically, the hybrid diversity value is lower than the pair-wise and non-pair-wise diversity values [28]. This distance is based on a reference h^*; therefore, this is exactly the distance between two clusterings. The pair-wise diversity value and hybrid diversity value, based on EJ criterion, are higher than the pair-wise diversity value and hybrid diversity value based on NMI. Therefore, the pair-wise and hybrid diversity based on EJ yielded higher quality value.

5.4. Test results based on quality

In this experiment, the effects of the proposed method and the proposed criterion on the quality of the result were investigated. For each dataset, 100 different base clusterings were produced. The result of the $d4$ measure, which was a hybrid measure based on EJ, was better than that of the other measures. The final cluster quality using different quality measures ($d1, d2, d3$, and $d4$) and full ensembles with different consensus functions,

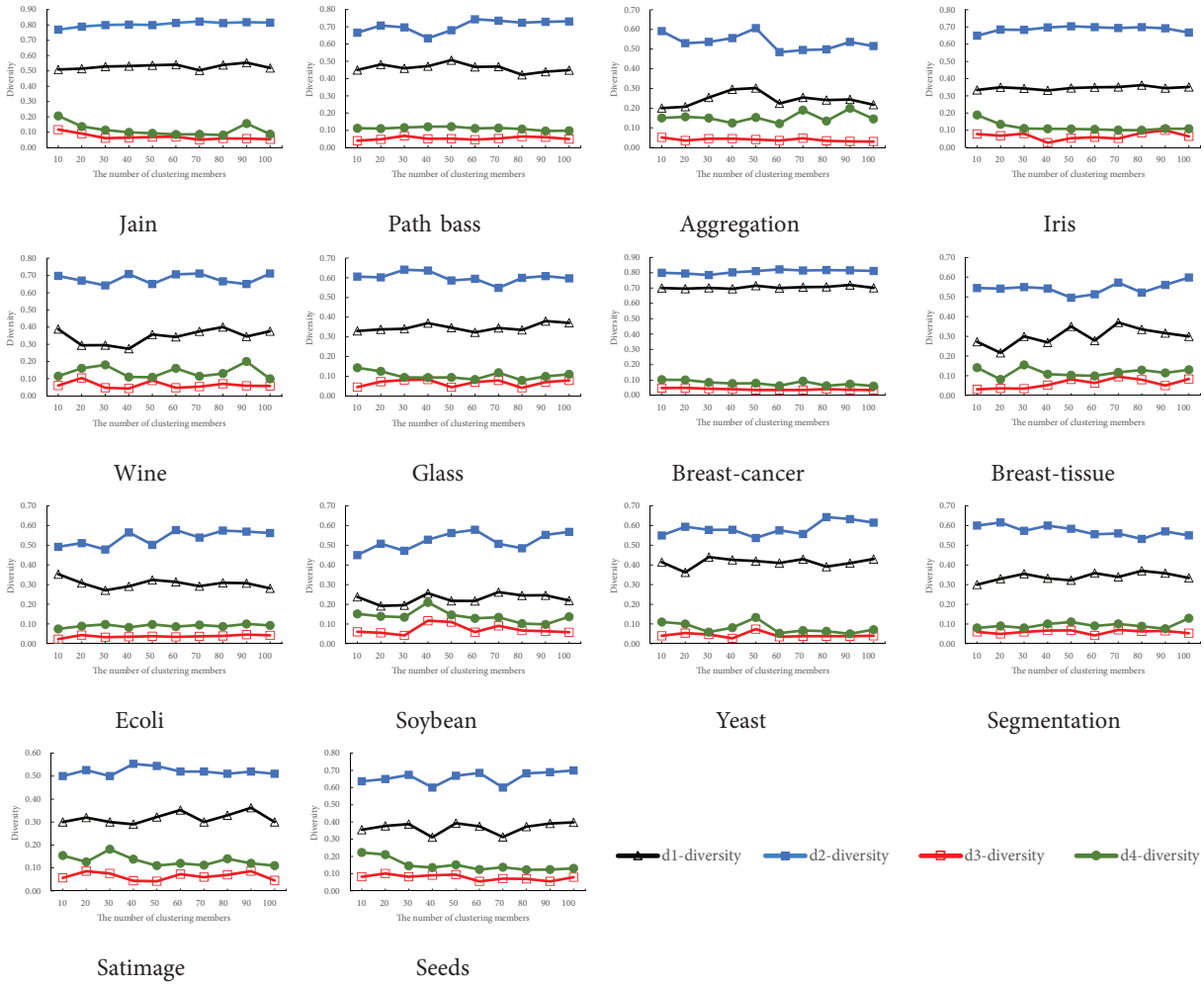

Figure 4. Comparing the diversity results obtained using four diversity measures with applying HGPA. Note that d1-diversity=$1 - NMI(h^i, h^j)$, d2-diversity=$1 - EJ(h^i, h^j)$, d3-diversity=$NMI(h^*, h^i) - NMI(h^*, h^j)$, and d4-diversity=$EJ(h^*, h^i) - EJ(h^*, h^j)$.

namely HGPA, CSPA, and MCLA, are displayed in Tables 5, 6, and 7. The maximum value for each dataset was bolded. As illustrated in Tables 5, 6, and 7, the $d4$ measure is better than the others.

6. Conclusion and future work

The present research was focused on offering an effective solution to the problem of cluster ensemble selection. Most of the existing CES algorithms have been designed on the basis of diversity and quality criteria. These algorithms typically have a tendency to assess the diversity and quality of primary clusterings by using the evaluation indices. Due to the fact that the number of subsets of base clusterings is uncertain, we made use of a hierarchical algorithm aiming to choose a subset comprising more effective clusters in a way to finally obtain a high-performance solution. In the proposed method, clustering partitions are chosen concerning their qualities, which are specified by the hierarchical method and new criterion. Three consensus functions, i.e., MCLA, HGPA, and CSPA, were employed to achieve the consensus solution and also to aggregate the selected subsets in the proposed method. The proposed method outperformed the full ensemble regarding the exploration of the

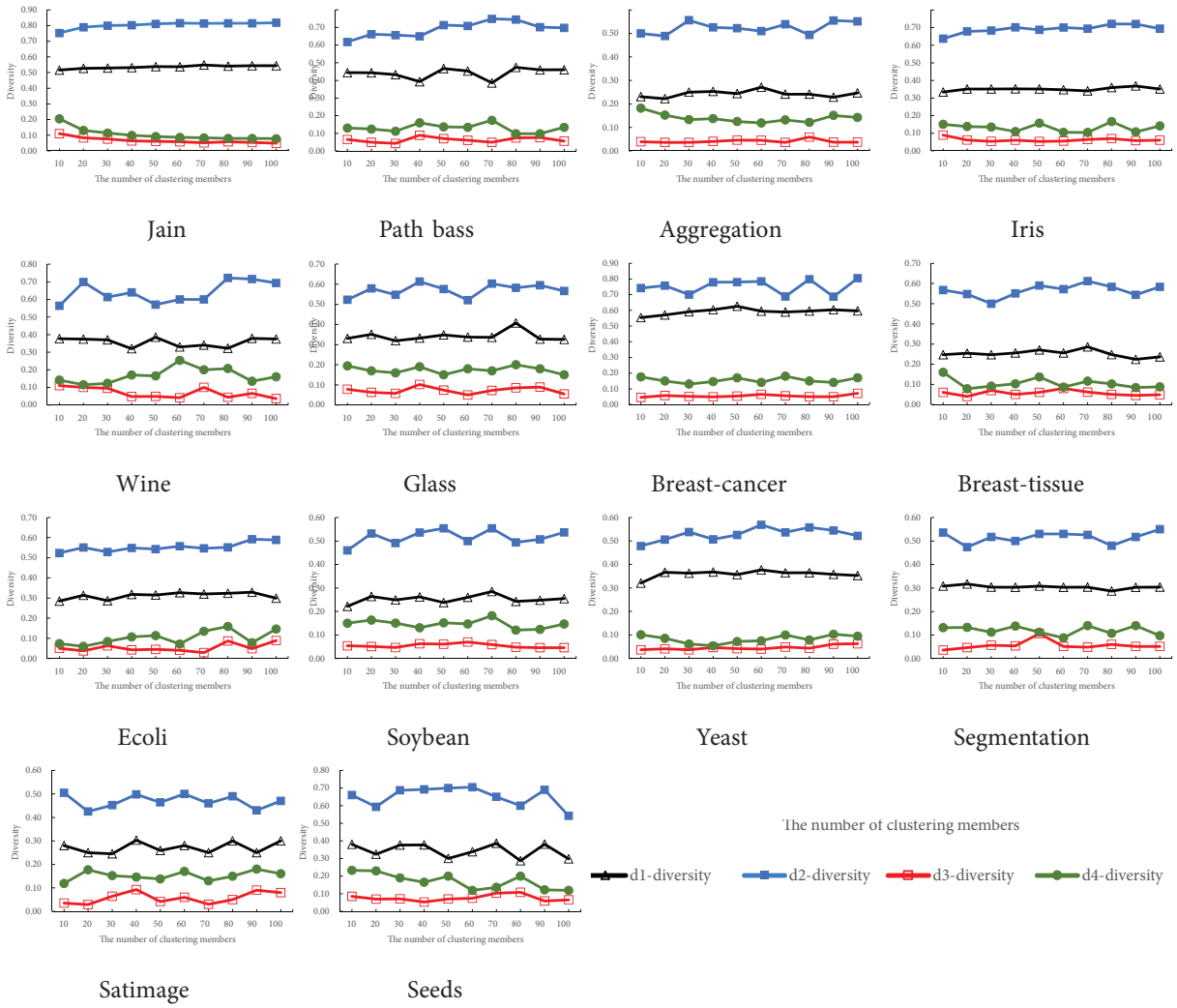

Figure 5. Comparing the diversity results obtained using four diversity measures with applying CSPA. Note that d1-diversity=$1 - NMI(h^i, h^j)$, d2-diversity=$1 - EJ(h^i, h^j)$, d3-diversity=$NMI(h^*, h^i) - NMI(h^*, h^j)$, and d4-diversity=$EJ(h^*, h^i) - EJ(h^*, h^j)$.

subset of cluster members on the basis of diversity and quality. The results of comparing the proposed algorithm performance with that of different measures and the full ensemble on 14 datasets indicated that the proposed algorithm with EJ could make higher diversity and get more accurate results. It was concluded that the use of the EJ criterion as diversity and quality measure obtained better results than the NMI criterion. It was also confirmed that the use of the hybrid diversity measure based on EJ was the best option for creating higher diversity. Future work in this field can be focused further on exploring the effects of noise and missing values of the data upon the EJ criterion and also on studying the application of the proposed method to different domains.

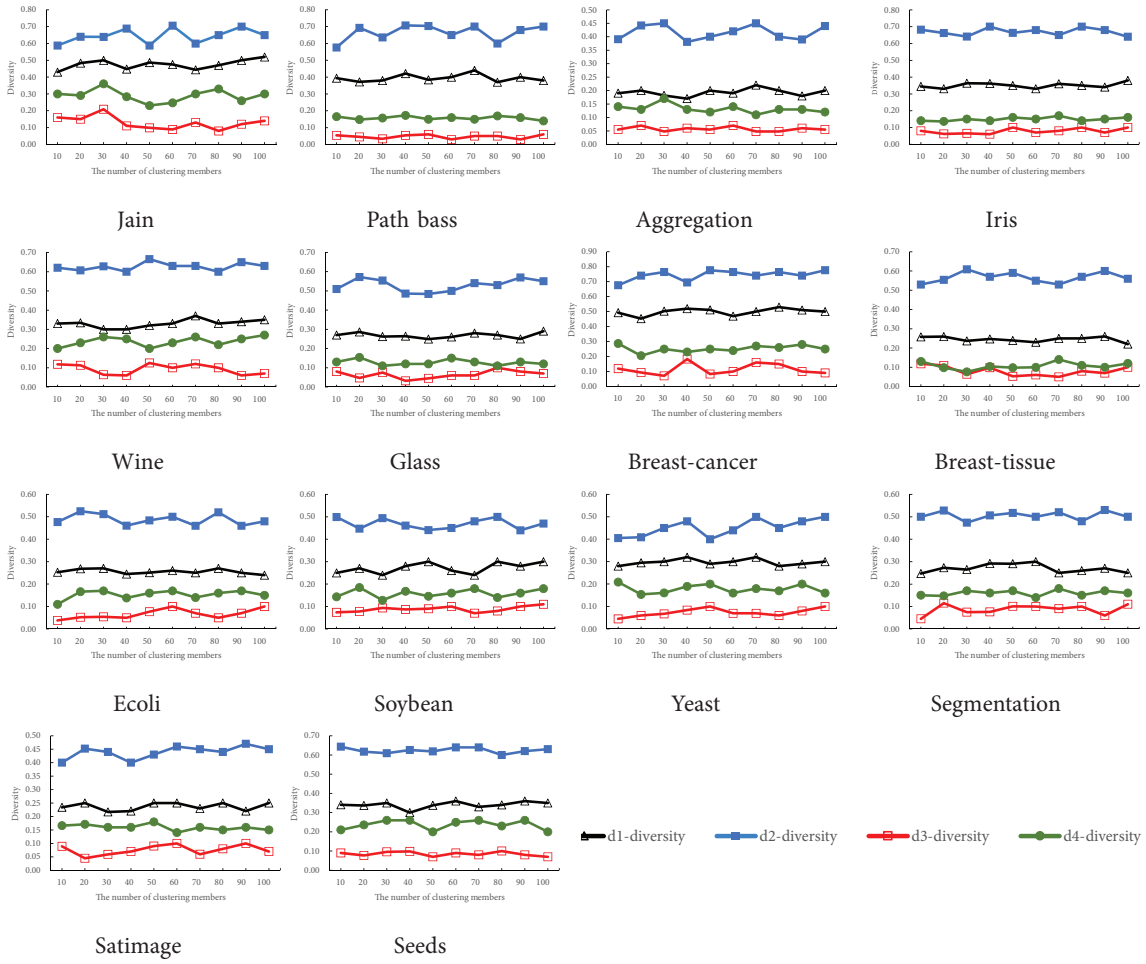

Figure 6. Comparing the diversity results obtained using four diversity measures with applying MCLA. Note that d1-diversity=$1 - NMI(h^i, h^j)$, d2-diversity=$1 - EJ(h^i, h^j)$, d3-diversity=$NMI(h^*, h^i) - NMI(h^*, h^j)$, and d4-diversity=$EJ(h^*, h^i) - EJ(h^*, h^j)$.

Table 5. Clustering quality using different quality measures and full ensembles with HGPA algorithm.

Dataset	Full ensembles	d1-accuracy	d2-accuracy	d3-accuracy	d4-accuracy
Jain	0.306	0.373	0.373	0.376	**0.422**
Path bass	0.459	0.769	0.798	0.788	**0.816**
Aggregation	0.791	0.802	0.802	**0.805**	**0.805**
Soybean(small)	0.679	0.884	**0.910**	**0.910**	0.777
Breast-tissue	0.338	0.341	0.355	0.367	**0.377**
Iris	0.901	0.901	0.901	**0.919**	**0.919**
Wine	0.425	0.437	0.437	0.437	**0.469**
Seeds	0.702	0.713	**0.732**	0.718	**0.732**
Glass	0.321	0.359	0.380	0.365	**0.416**
Ecoli	0.500	0.561	0.565	0.525	**0.580**
Breast-cancer	0.485	0.485	0.485	**0.499**	**0.499**
Yeast	0.199	0.232	**0.236**	0.227	0.230
Segmentation	0.334	0.530	0.524	0.541	**0.566**
Satimage	0.473	0.492	0.585	0.512	**0.595**

Table 6. Clustering quality using different quality measures and full ensembles with CSPA algorithm.

Dataset	Full ensembles	d1-accuracy	d2-accuracy	d3-accuracy	d4-accuracy
Jain	**0.360**	**0.360**	**0.360**	**0.360**	**0.360**
Path bass	0.575	0.753	**0.786**	0.661	0.663
Aggregation	0.733	0.733	**0.737**	**0.737**	**0.737**
Soybean(small)	0.754	0.802	0.822	0.842	**0.853**
Breast-tissue	0.315	0.366	0.378	0.380	**0.383**
Iris	0.919	**0.939**	**0.939**	0.918	0.925
Wine	0.346	0.393	0.452	0.471	**0.478**
Seeds	0.684	0.726	**0.735**	0.685	0.721
Glass	0.367	0.389	0.402	0.390	**0.422**
Ecoli	0.513	0.528	0.535	0.536	**0.543**
Breast-cancer	0.485	**0.501**	**0.501**	0.485	0.485
Yeast	0.227	0.235	0.241	0.238	**0.246**
Segmentation	0.488	0.494	0.510	0.500	**0.552**
Satimage	0.551	0.553	0.550	0.555	**0.577**

Table 7. Clustering quality using different quality measures and full ensembles with MCLA algorithm.

Dataset	Full ensembles	d1-accuracy	d2-accuracy	d3-accuracy	d4-accuracy
Jain	0.366	0.366	0.416	0.373	**0.421**
Path bass	0.540	0.572	0.586	0.589	**0.605**
Aggregation	0.790	0.870	0.882	**0.896**	0.890
Soybean(small)	0.862	**1.000**	**1.000**	**1.000**	**1.000**
Breast-tissue	0.368	0.382	**0.415**	0.403	**0.415**
Iris	**0.919**	**0.919**	**0.919**	**0.919**	**0.919**
Wine	0.395	0.452	0.457	0.467	**0.474**
Seeds	0.695	0.746	**0.762**	0.739	0.720
Glass	0.345	0.416	0.418	**0.433**	0.420
Ecoli	0.587	0.606	0.617	**0.622**	**0.622**
Breast-cancer	0.812	0.822	0.836	0.831	**0.848**
Yeast	0.243	0.259	0.265	0.257	**0.268**
Segmentation	0.549	0.609	0.592	0.598	**0.632**
Satimage	0.524	0.592	**0.620**	0.612	**0.620**

References

[1] Jain AK. Data clustering: 50 years beyond k-means. Pattern Recognition Letters 2010; 31 (8): 651-666. doi: 10.1016/j.patrec.2009.09.011

[2] Alizadeh H, Minaei-Bidgoli B, Parvin H. Cluster ensemble selection based on a new cluster stability measure. Intelligent Data Analysis 2014; 18 (3): 389-408. doi: 10.3233/IDA-140647

[3] Yu Z, Chen H, You J, Han G, Li L. Hybrid fuzzy cluster ensemble framework for tumor clustering from biomolecular data. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2013; 10 (3): 657-670. doi: 10.1109/TCBB.2013.59

[4] Sun J, Chen W, Fang W, Wun X, Xu W. Gene expression data analysis with the clustering method based on an improved quantum-behaved Particle Swarm Optimization. Engineering Applications of Artificial Intelligence 2012; 25 (2): 376-391. doi: 10.1016/j.engappai.2011.09.017

[5] Wu X, Ma T, Cao J, Tian Y, Alabdulkarim A. A comparative study of clustering ensemble algorithms. Computers & Electrical Engineering 2018; 68: 603-615. doi: 10.1016/j.compeleceng.2018.05.005

[6] Strehl A, Ghosh J. Cluster ensembles–a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research 2002; 3 (Dec): 583-617. doi: 10.1162/153244303321897735

[7] Li F, Qian Y, Wang J, Liang J. Multigranulation information fusion: A Dempster-Shafer evidence theory-based clustering ensemble method. Information Sciences 2017; 378: 389-409. doi: 10.1016/j.ins.2016.10.008

[8] Kuncheva LI, Hadjitodorov ST. Using diversity in cluster ensembles. In: IEEE 2004 International Conference on Systems, Man and Cybernetics; The Hague, Netherlands; 2004. pp. 1214-1219. doi: 10.1109/ICSMC.2004.1399790

[9] Yang F, Li X, Li Q, Li T. Exploring the diversity in cluster ensemble generation: Random sampling and random projection. Expert Systems with Applications 2014; 41 (10): 4844-4866. doi: 10.1016/j.eswa.2014.01.028

[10] Hamidi SS, Akbari E, Motameni H. Consensus clustering algorithm based on the automatic partitioning similarity graph. Data & Knowledge Engineering 2019; 124: 101754. doi: 10.1016/j.datak.2019.101754

[11] Vega-Pons S, Ruiz-Shulcloper J. A survey of clustering ensemble algorithms. International Journal of Pattern Recognition and Artificial Intelligence 2011; 25 (03): 337-372. doi: 10.1142/S0218001411008683

[12] Azimi J, Fern X. Adaptive cluster ensemble selection. In: Proceedings of the 21st international joint conference on Artificial intelligence; San Francisco, CA, USA; 2009. pp. 992-997.

[13] Hong Y, Kwong S, Wang H, Ren Q. Resampling-based selective clustering ensembles. Pattern Recognition Letters 2009; 30 (3): 298-305. doi: 10.1016/j.patrec.2008.10.007

[14] Hadjitodorov ST, Kuncheva LI, Todorova LP. Moderate diversity for better cluster ensembles. Information Fusion 2006; 7 (3): 264-275. doi: 10.1016/j.inffus.2005.01.008

[15] Fern XZ, Lin W. Cluster ensemble selection. Statistical Analysis and Data Mining: The ASA Data Science Journal 2008; 1 (3): 128-141. doi: 10.1002/sam.10008

[16] Wang X, Han D, Han C. Rough set based cluster ensemble selection. In: Proceedings of the 16th International Conference on Information Fusion; İstanbul, Turkey; 2013. pp. 438-444.

[17] Ren Y, Domeniconi C, Zhang G, Yu G. Weighted-object ensemble clustering: methods and analysis. Knowledge and Information Systems 2017; 51 (2): 661-689. doi: 10.1007/s10115-016-0988-y

[18] Huang D, Wang CD, Lai JH. Locally Weighted Ensemble Clustering. IEEE transactions on cybernetics 2018; 48 (5): 1460-1473. doi: 10.1109/TCYB.2017.2702343

[19] Parvin H, Minaei-Bidgoli B, Alizadeh H. A new clustering algorithm with the convergence proof. In: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems; Berlin, Heidelberg, Germay; 2011. pp. 21-31. doi: 10.1007/978-3-642-23851-2_3

[20] Saidi M, Bechar ME, Settouti N, Chikh MA. Instances selection algorithm by ensemble margin. Journal of Experimental & Theoretical Artificial Intelligence 2018; 30 (3): 457-478. doi: 10.1080/0952813X.2017.1409283

[21] Lu X, Yang Y, Wang H. Selective clustering ensemble based on covariance. In: International Workshop on Multiple Classifier Systems; Berlin, Heidelberg, Germany; 2013. pp. 179-189. doi: 10.1007/978-3-642-38067-9_16

[22] Naldi MC, Carvalho AC, Campello RJ. Cluster ensemble selection based on relative validity indexes. Data Mining and Knowledge Discovery 2013; 27 (2): 259-289. doi: 10.1007/s10618-012-0290-x

[23] Hubert L, Arabie P. Comparing partitions. Journal of classification 1985; 2 (1): 193-218. doi: 10.1007/BF01908075

[24] Zhong S, Ghosh J. Generative model-based document clustering: a comparative study. Knowledge and Information Systems 2005; 8 (3): 374-384. doi: 10.1007/s10115-004-0194-1

[25] Kandylas V, Upham SP, Ungar LH. Finding cohesive clusters for analyzing knowledge communities. Knowledge and Information Systems 2008; 17 (3): 335-354. doi: 10.1007/s10115-008-0135-5

[26] Alizadeh H, Minaei-Bidgoli B, Parvin H. To improve the quality of cluster ensembles by selecting a subset of base clusters. Journal of Experimental & Theoretical Artificial Intelligence 2014; 26 (1): 127-150. doi: 10.1080/0952813X.2013.813974

[27] Jia J, Xiao X, Liu B, Jiao L. Bagging-based spectral clustering ensemble selection. Pattern Recognition Letters 2011; 32 (10): 1456-1467. doi: 10.1016/j.patrec.2011.04.008

[28] Akbari E, Dahlan HM, Ibrahim R, Alizadeh H. Hierarchical cluster ensemble selection. Engineering Applications of Artificial Intelligence 2015; 39: 146-156. doi: 10.1016/j.engappai.2014.12.005

[29] Yousefnezhad M, Reihanian A, Zhang D, Minaei-Bidgoli B. A new selection strategy for selective cluster ensemble based on diversity and independency. Engineering Applications of Artificial Intelligence 2016; 56: 260-272. doi: 10.1016/j.engappai.2016.10.005

[30] Yang F, Li T, Zhou Q, Xiao H. Cluster ensemble selection with constraints. Neurocomputing 2017; 235 (1): 59-70. doi: 10.1016/j.neucom.2017.01.001

[31] Shi Y, Yu Z, Chen CP, You J, Wong HS, Wang Y, Zhang J. Transfer clustering ensemble selection. IEEE Transactions On Cybernetics 2018; 50 (6): 2872-2885. doi: 10.1109/TCYB.2018.2885585

[32] Wang H, Liu G. Two-level-oriented selective clustering ensemble based on hybrid multi-modal metrics. IEEE Access 2018; 6: 64159-64168. doi: 10.1109/ACCESS.2018.2877666

[33] Li F, Qian Y, Wang J, Dang C, Liu B. Cluster's quality evaluation and selective clustering ensemble. ACM Transactions on Knowledge Discovery from Data (TKDD) 2018; 12 (5): 1-27. doi: 10.1145/3211872

[34] Abbasi SO, Nejatian S, Parvin H, Rezaie V, Bagherifard K. Clustering ensemble selection considering quality and diversity. Artificial Intelligence Review 2019; 52 (2): 1311-1340. doi: 10.1007/s10462-018-9642-2

[35] Liang W, Zhang Y, Xu J, Lin D. Optimization of basic clustering for ensemble clustering: an information-theoretic perspective. IEEE Access 2019; 7: 179048-179062. doi: 10.1109/ACCESS.2019.2950159

[36] Sibson R. SLINK: an optimally efficient algorithm for the single-link cluster method. The Computer Journal 1973; 16 (1): 30-34. doi: 10.1093/comjnl/16.1.30